UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Tema 14

Control del riesgo alfa

Diseños adaptativos

Jordi Cortés y Erik Cobo

Héctor Rufino, Marta Vilaró y José Antonio González

2014

Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

equator network

UPC

ELSEVIER DOYMA

MEDICINA CLINICA

TRIALS

Control del riesgo alfa

Presentación

Este capítulo aborda como adaptar el riesgo α a las necesidades del estudio al mismo tiempo que garantiza que a nivel global no supera el límite deseado –usualmente un 5%.

La primera parte, multiplicidad, expone el problema y diferentes soluciones generales.

La segunda parte explica los diseños que permiten *adaptar* el reclutamiento, tamaño muestral, criterios de inclusión, variable principal, o la razón de asignación a los tratamientos —por ejemplo, pueden ser modificados durante su ejecución dependiendo de los resultados obtenidos en el análisis. Por supuesto, debe especificarse así en el protocolo, ya que de lo contrario el *diseño* no sería adaptativo.

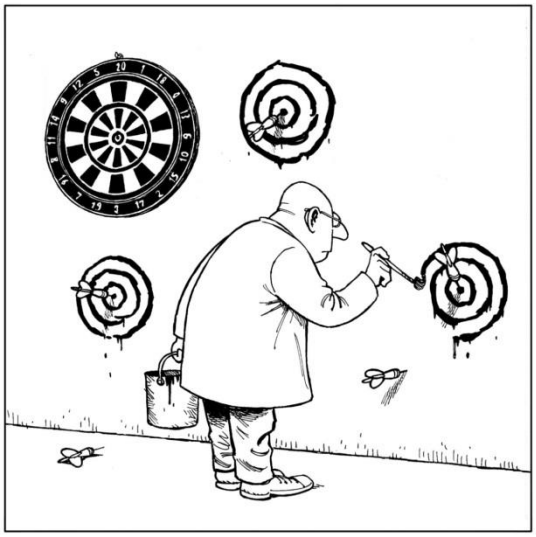
La tercera parte aborda los ensayos clínicos secuenciales, por ser los adaptativos más frecuentes y mejor aceptados por las agencias reguladoras. En esta clase de diseños, la adaptabilidad hace referencia al tamaño muestral, ya que éste dependerá de los resultados obtenidos en análisis intermedios. Los datos son analizados en determinados instantes pre-establecidos con el fin de demostrar la eficacia del tratamiento o la futilidad del diseño para establecerlo. De esta forma, se actualiza la información hipotética usada en el cálculo del tamaño muestral y se puede alcanzar una conclusión en el momento justo, resultando en tamaños muestrales menores que los diseños clásicos de muestra fija y, por consiguiente, a un coste humano y económico menor, al mismo tiempo que se agiliza el acceso de los pacientes a las nuevas intervenciones (parada por eficacia) o se acorta su innecesaria exposición en estudios previsiblemente 'negativos' (parada por futilidad).

Contribuciones: Basados en transparencias de Erik Cobo y José Antonio González; Jordi Cortés preparó la parte de "diseños adaptativos y secuenciales"; y Hector Rufino y Erik Cobo la de "Multiplicidad"; que han sido revisadas todas ellas por Marta Vilaró y Erik Cobo.

1. Multiplicidad

Los riesgos deben gestionarse con prudencia. Hemos aceptado que un estudio asuma un pequeño riesgo de autorizar una intervención no eficaz, una vez. Pero si este proceso se repite indefinidamente, sin duda se cometerá este error.

Nota: Tanto va el cántaro a la fuente que al final se rompe.



Recuerde

No abuse de las pruebas de hipótesis.

Ejercicio 1.1

Los EC pivote pretenden una decisión, sea cambiar la guía clínica habitual, sea registrar una nueva intervención. ¿Puede tener 'k' variables principales y una variable secundaria?

Recuerde

El protocolo de un EC pivote, si aumenta las variables o las pruebas, debe definir las reglas de decisión.

Ejemplo 1.1: Para estudiar el efecto de T frente a C sobre tres respuestas de interés Y_1, Y_2, Y_3, se hacen tres contrastes, cada uno con un riesgo $\alpha = 0.05$. La regla de decisión podría funcionar por intersección (positivo si las 3 pruebas fueran positivas) o por unión (positivo si cualquiera fuera positiva). Es decir, en

el primer caso, se autorizaría el producto sólo si las 3 pruebas resultan significativas; en el segundo, bastaría con que lo fuera una de ellas. En el primer caso, el estudio pierde potencia (hay menos opciones de alcanzar el mercado de las que tendría una sola prueba); en el segundo, sobre-consume α (cada contraste "gasta" α).

Nota: Asumiendo, por simplicidad, que las 3 pruebas son independientes, se puede calcular la pérdida de potencia si el proceso exige que las 3 pruebas sean significativas. Tomando una potencia del 85% para cada prueba i, que equivale a un β_i=0.15, la potencia global es:

$$1 - \beta_G = (1 - \beta_i)^3 = (1 - 0.15)^3 = 0.614$$

Es decir, que si la intervención tuviera el efecto especificado, las probabilidades de fallar al intentar demostrarlo serían 0.386, ¡casi un 40%! Ningún promotor querría fallar en 4 de cada 10 intervenciones eficaces.

Recuerde

Perderá potencia si requiere que todas las pruebas sean significativas.

En el segundo caso, en cambio, si el criterio para autorizar la intervención solo requiere que una de las 3 pruebas fuera significativa, se pierde el control del riesgo α y la probabilidad de autorizar una intervención no eficaz es mayor del 5%, ya que asumimos este riesgo en 3 ocasiones.

Nota: Asumiendo otra vez independencia entre los resultados de las 3 pruebas y considerando un α=0.05 para cada una, el error de tipo I global sería también mayor de lo deseado:

$$\alpha_G = 1 - (1 - 0.05)^3 = 0.14$$

Es decir, que si la intervención no tuviera efecto, un 14% de estudios conducirían a administrarlas: ninguna agencia de regulación aceptaría que 1 de cada 7 intervenciones no eficaces terminara siendo aconsejada a los pacientes.

Así, cuando basta que una de las pruebas sea significativa para considerar el estudio positivo, se pierde el control del riesgo α y se habla de multiplicidad.

Recuerde

Gasta, consume o pierde el control del riesgo α si realiza múltiples pruebas y se queda con la significativa.

La necesidad de ajustar por multiplicidad ha sido muy discutida.

Ejemplo 1.2. Shulz imagina un estudio con 2 respuestas relacionadas: 50% de reducción de fiebre (RR=0.5, $IC_{95\%}$ de 0.25 a 0.99, *P*=0.041) y 52% de reducción de infección (RR=0.48, $IC_{95\%}$ de 0.24 a 0.97, *P*=0.041). Aquellos contrarios al ajuste opinarían que ambos resultados positivos se apoyan mutuamente, mientras que los favorables al ajuste dirían que el consumo global de α supera el 5% y, por haber hecho 2 pruebas, los resultados no son significativos.

Ejercicio 1.2

¿Resuelve esta ambigüedad especificar en el protocolo el criterio de decisión?

1.1. Objetivo del EC

Hay que diferenciar si el objetivo es hacer inferencia o tomar una decisión. Si el propósito de realizar diferentes pruebas de hipótesis es hacer inferencia sobre varias preguntas de conocimiento, es usual argumentar que son preguntas diferentes y que no tiene sentido considerar a las diferentes pruebas parte de un objetivo común.

Ejemplo 1.3: Un investigador puede estar interesado en conocer sobre qué variables de respuesta (presión arterial sistólica, diastólica, media, diferencial, a la semana, al mes, al trimestre, etc.) hasta un total de 10 se manifiesta el efecto de una intervención. Como cada prueba contesta una pregunta diferente, se puede argumentar que no es necesario ajustar por multiplicidad.

Así, las revistas científicas no tienen una postura clara sobre la conveniencia de ajustar por multiplicidad. En cambio, si la intención es tomar una decisión única en base a todas las pruebas, el error α debe calcularse considerando las diversas opciones que tiene el estudio de alcanzar su objetivo.

Ejemplo 1.4: Un promotor quiere comparar un nuevo tratamiento con el control sobre los 10 indicadores anteriores. El objetivo es sacar al mercado el nuevo tratamiento, si su efecto es significativo en alguno de los diez indicadores. Se toma un nivel de significación individual α_i=5%, se obtiene un intolerable nivel de significación global α_G:

$$\alpha_G = 1 - (1 - 0.05)^{10} = 0.599!!$$

Las agencias de regulación de intervenciones sanitarias tienen una postura muy clara.

Recuerde

La multiplicidad se define bien en el entorno de decisión.

1.2. Hipótesis frente a premisas

En ocasiones, se utilizan pruebas de *hipótesis* para estudiar *premisas*. Las guías de publicación dicen claramente que conviene concentrar los riesgos estadísticos en los objetivos del estudio. La pregunta de si las premisas son ciertas es secundaria. Más interesante es un análisis de sensibilidad que permita saber si, bajo otras premisas, se llega a la misma conclusión

Ejemplo 1.5: Buysé et al muestran que sus conclusiones son las mismas sea cual sea el punto de corte que escogen para la variable respuesta.

Ejercicio 1.3

STROBE E&E 12e dice:

a) Hay que poner a prueba las premisas en las que descansa el estudio y su análisis (como la Normalidad de la respuesta)
b) Conviene hacer análisis de sensibilidad para ver hasta qué punto las conclusiones son consistentes o bien dependen de las premisas.
c) No dice nada.

Ejercicio 1.4

¿Cuáles de los siguientes dice STROBE E&E 12e que puede abarcar el análisis de sensibilidad?

a) Criterios de inclusion en los análisis
b) Definición de la exposición
c) Definición de las respuestas
d) Tratamiento de los datos ausentes
e) Sesgos introducidos por el proceso de medida
f) Elecciones concretas en el análisis, como el tratamiento de las variables cuantitativas.

1.3. Error global (*Family Wise Error o FWE*)

Para poder distinguir entre error individual y global, lo primero que hay que definir es qué abarca el término 'global'. Para ello, se define a la familia de k pruebas de significación:

$$\{H\} = \{H_{01}, H_{02}, \dots, H_{0k}\}$$

como el conjunto de contrastes que, en caso de resultar significativos a nivel individual, permitirían tomar la decisión de interés.

Definición

El riesgo α_G global es la probabilidad de adoptar la decisión alternativa por rechazar al menos una hipótesis nula de la familia {H} siendo todas ellas ciertas.

A diferencia del riesgo individual α_i, que hace referencia a la prueba i, el α_G se interpreta como el riesgo global, acumulado para las k comparaciones.

1.4. Control disminuyendo el riesgo individual

Para obtener un riesgo α_G global igual o cercano al valor deseado (normalmente del 5%), la primera estrategia es disminuir el riesgo α_i individual.

1.4.1. Método de Bonferroni

La desigualdad de Boole establece que la probabilidad de que ocurra algún evento es como mucho igual a la suma de las probabilidades de todos los eventos considerados.

Ejemplo 1.6: Si accidente cardiovascular (AVC) incluye infarto de miocardio (IM), ictus (I) y accidente vascular periférico (AVP), dado que algunos casos presentan simultáneamente más de uno, la probabilidad de tener algún AVC es como mucho la suma de las probabilidades de IM, I y AVP: P(AVC) ≤ P(IM)+P(I)+P(AVP)

Así, la desigualdad de Boole establece que el riesgo α_G global será, como mucho, la suma de los riesgos asumidos en todos los contrastes. El método de Bonferroni propone repartir el riesgo α_G global entre todos los contrastes de forma que la suma de los

riesgos α_i individuales iguale al riesgo α_G global deseado. Si considerar igual a todas las hipótesis, asigna el mismo riesgo α_i individual a cada contraste.

Definición

Para garantizar $\alpha_G = \alpha$ con k contrastes, Bonferroni $\alpha_i = \alpha/k$.

Ejemplo 1.7: Se desea comparar el efecto de 3 nuevos tratamientos para el cáncer de Mama frente al tratamiento convencional, con el objetivo de sacar al mercado el tratamiento (de los 3) que resulte significativo. Si quiere tener un riesgo α_G global=0.05, el $\alpha_i$individual será:

$$\alpha_i = \frac{0.05}{3} = 0.0167$$

Nota: Se trata de una desigualdad: por lo general, el riesgo global será inferior a la suma de los riesgos individuales: se garantiza que no supera el riesgo $\boldsymbol{\alpha}$ global deseado (¡bien!), pero se podría estar perdiendo más potencia de la necesaria (¡mal!).

Ejercicio 1.5

¿Cuál debería ser el riesgo individual α_i si quiere aplicar el método de Bonferroni en un EC pivote con 10 variables respuesta principales y se desea mantener el riesgo α global α_G=0.05? Interprete.

1.4.2. Método de Sidák

Nota: Al inicio hemos ilustrado el problema con este método.

Igual que el anterior, Sidák desciende el riesgo α_i individual para obtener un riesgo α_G global deseado, pero ahora asume independencia entre las pruebas realizadas para poder multiplicar sus probabilidades.

Definición

Para garantizar $\alpha_G = \alpha$ con k pruebas, Sidák $\alpha_i = 1 - (1 - FWE)^{1/k}$.

Ejemplo 1.8: Siguiendo con el ejemplo anterior, si se deseara utilizar el método de Sidák para controlar la multiplicidad y garantizar α_G = 0.05:

$$\alpha_i = 1 - (1 - 0.05)^{\frac{1}{3}} = 0.0169$$

Valor muy similar al obtenido por Bonferroni (0.0167).

Ejercicio 1.6

a) Idem ejercicio 1.4 para Sidák.

b) ¿Cree que estas 2 estrategias tienen algún *efecto colateral*?

Nota: Bonferroni y Sidák dan resultados similares si k y α son pequeños (demostración por series de Taylor).

Recuerde

Disminuir el riesgo α_i de la prueba iésima disminuye también la potencia de esta prueba.

1.5. Grado de nulidad de la hipótesis

En una familia de k pruebas de hipótesis conviene valorar si las conclusiones de un contraste tienen implicaciones sobre los otros.

Definición

En una combinación restringida, el rechazo de un contraste implica cambios en otros.

Ejemplo 1.9: Sean 3 intervenciones, A, B y C; y 3 hipótesis que se desean contrastar: H_{01}: $\mu_A = \mu_B$; H_{02}: $\mu_A = \mu_C$; y H_{03}: $\mu_B = \mu_C$. Si rechazamos H_{01}: $\mu_A = \mu_B$, entonces H_{02}: $\mu_A = \mu_C$ y H_{03}: $\mu_B = \mu_C$ no pueden ser ambas ciertas.

Como para cometer el riesgo α es necesario que H_0 sea cierta, sólo hay que controlar la multiplicidad para el conjunto de Hipótesis que pueden ser simultáneamente ciertas.

1.6. Rechazo secuencial de hipótesis

Una vez se ha rechazado cierta H_{0i} ya no tiene sentido seguir asumiendo que es cierta y, por tanto, no es necesario protegerla ante multiplicidad.

Definición

Holm ordena los *P* valores de más a menos significativos y los pone a prueba sucesivamente ajustando (Bonferroni) cada uno sólo por las hipótesis aún no rechazadas.

Ejemplo 1.10: Los 5 valores de P observados han sido: 0.0021, 0.0093, 0.0137, 0.0324 y 0.1188. Al contrastar el primero debe controlar que hasta k=5 hipótesis nulas podrían ser ciertas, por lo que $\alpha_i = \alpha_G/\mathrm{k} = 0.05/5 = 0.01$ >P=0.0021, se rechaza H_{01}. Pero para contrastar la segunda H_{02}, ya no es necesario protegerse por si H_{01} fuera cierta, por lo que K=4 y $\alpha_i = \alpha_G/\mathrm{k} = 0.05/4 = 0.0125$ >P=0.0093 también se rechaza H_{012}.

Ejercicio 1.7

Termine el proceso de Holm para las 3 siguientes pruebas.

Nota: Shaffer perfiló el método de Holm ajustado sólo por las restantes pruebas que podrían ser simultáneamente ciertas.

Definición

Hockberg ordena los P valores al revés, de menos a más significativos y los contrasta sucesivamente ajustando (Bonferroni) cada uno sólo por las hipótesis previamente no rechazadas.

Ejemplo 1.11: con los mismos 5 valores de P anteriores, el primero que se contrasta ahora es 0.1188, que al ser mayor que 0.05, no se rechaza. Al mirar el segundo hay que tener en cuenta que 2 podrían ser simultáneamente ciertos, por lo que 0.0324>0.05/2=0.025, tampoco se rechaza.

Ejercicio 1.8

Termine el proceso de Hockberg para las otras 3 pruebas.

Con el paquete de R *multtest* se pueden realizar pruebas de hipótesis múltiples utilizando los diferentes métodos de ajuste vistos en este tema. En concreto, la función *mt.rawp2adjp* devuelve los P valores ajustados para los diferentes métodos.

Ejemplo de R

```
# Instalación y carga de 'multtest'
> source("http://bioconductor.org/biocLite.R")
> biocLite("multtest")
> library(multtest)
```

```
# Aplicación al Ejemplo 1.10
#Creamos un vector que contenga los P valores obtenidos
> P <- c(0.0021,0.0093,0.0137,0.0324,0.1188)
#Indicamos los prodecimientos que queremos utilizar
> procs <- c("Bonferroni", "Holm", "Hochberg","SidakSS")
#Utilizamos la función mt.rawp2adjp
> res <- mt.rawp2adjp(P, procs)
> adjp <- res$adjp[order(res$index), ]
> round(adjp,3)
     rawp Bonferroni  Holm Hochberg SidakSS
[1,] 0.002      0.010 0.010    0.010   0.010
[2,] 0.009      0.046 0.037    0.037   0.046
[3,] 0.014      0.068 0.041    0.041   0.067
[4,] 0.032      0.162 0.065    0.065   0.152
[5,] 0.119      0.594 0.119    0.119   0.469
#Ahora, se pueden comparar los p valores ajustados de cada método con el 5%, para ver si se acepta o se rechaza la hipótesis nula.
#R hace el cálculo con toda la precisión, por lo que los valores redondeados que proporciona pueden no cuadrar. P.e.: 0.002*5=0.010, pero 0.009*5=0.045≈0.046.
```

Ejercicio 1.9

¿Por qué cambia la conclusión para la prueba "[3,] 0.014" de los 4 métodos anteriores?. ¿Por qué coinciden 2 a 2?

Ejercicio 1.10

Se ha realizado un ECA para estudiar el efecto de los hábitos higiénicos (ejercicio, dieta, siesta,...) en 7 variables de constantes vitales y lipemias obteniendo: PAS P= 0.012; PAD P= 0.011; FC P=0.467; HDL P=0.006; LDL P=0.314; CT P=0.123; y T P=0.08. Realice un ajuste por multiplicidad para un α_G global de 0.05, según los métodos de (a) Bonferroni, (b) Sidak, (c) Holm (+Bonferroni) y (d) Hochberg (+Bonferroni).

Nota: los métodos de Newman-Keuls y de Duncan son aplicaciones del método secuencial a las comparaciones entre k grupos.

1.7. Método de pruebas cerradas bajo intersección*

El principio de pruebas cerradas bajo intersección establece que no es necesario ajustar una hipótesis por multiplicidad si está contenida en la hipótesis previamente rechazada.

Ejemplo 1.12: Bajo la premisa de efecto no decreciente dentro del rango de dosis estudiado, afirmar que la dosis de 3 g iguala a la dosis 0 g (H_{03}), implica que también las dosis 2 g y 1 g igualan a la dosis 0 g (H_{02} y H_{01}). Así, se pone primero a prueba H_{03} y sólo si se rechaza se sigue con H_{02} y, si también se rechaza, con H_{01}. Como bajo la premisa de efecto no decreciente, si 3 g, no tiene efecto, tampoco lo tiene 2 g, H_{02} está contenida en H_{03}. Poner H_{02} a prueba sólo si H_{03} ha sido rechazada implica que el riesgo de H_{02} está dentro del de H_{03} y no es necesario realizar ajustes: todas ellas se ponen a prueba con α=0.05.

Ejercicio reto

Un fármaco ha sido probado a dosis de 0, 1, 2, 3, 4, 5 y 6 mg/Kg en 7 subgrupos de 3 casos cada uno, habiéndose obtenido las respuestas medias 12.88, 12.86, 12.82, 14.12, 14.08, 13.99 y 14.00, con una desviación típica común intragrupo (*pooled:* S_P) de 0,617. El límite de significación de tablas es $t_{14,0.975}$=2.145 (ya que S_P está estimada con 14 gdl). Calcule el estadístico t (señal/ruido) para todas la comparaciones respecto a la dosis de 0 mgr y responda qué dosis son distintas de la de 0mg/Kg bajo el principio de pruebas cerradas bajo intersección.

1.8. Pruebas fisherianas y métodos de remuestreo*

Fisher dijo que, si la hipótesis nula fuera cierta, cualquier asignación posible bajo el esquema de aleatorización tenía una probabilidad cuantificable de ser observada.

Ejemplo 1.13: Si asignamos al azar a los pacientes 1, 2, 3 y 4 a dos intervenciones T y C de forma que 2 sean asignados a cada una, las 6 posibles combinaciones TTCC, TCCT, TCTC, CTTC, CTCT y CCTT tienen todas ellas la misma probabilidad de ser observadas. Como hay 6 combinaciones de 4 elementos tomados de 2 en 2 , $\binom{4}{2} = \frac{4!}{2!2!} = \frac{4\cdot3\cdot2}{2\cdot2} = 6$, entonces cada una tiene una probabilidad igual a 1/6.

Ejemplo de R

```
# La función choose devuelve el número de combinaciones posibles.
> choose(4,2)
[1] 6
```

Ejercicio 1.11

¿Cuántas asignaciones posibles hay si queremos repartir 8 pacientes en dos grupos de forma equilibrada? ¿Qué probabilidad hay de que la asignación observada haya sido TTTTCCCC?

Ejemplo 1.14: La PAS de los 4 pacientes asignados a C ha sido 150, 147, 143 y 140; y la de los 4 asignados a T 130, 127, 123 y 120. Las medias respectivas son 145 y 125, con una estimación del efecto de 20 mmHg. Bajo la H_0 de T=C, este reparto tiene la misma probabilidad que cualquier otro, es decir, 1/70≈0.01428. Es decir, la probabilidad de que, por azar, los 4 pacientes asignados a la intervención T fueran los 4 de presión más baja es sólo de 0.01428. Cualquier otra asignación resultaría en una estimación menor del efecto. Por ello, si ordenamos todas las posibles asignaciones según la magnitud del efecto observado, vemos que cualquier otra asignación obtendría una estimación del efecto menor, por lo que el valor de P o "probabilidad de observar un valor como el observado o más extremo si asumimos cierta H_0" es, precisamente, 0.01428. Como es menor que 0.05, se rechaza H_0.

Las pruebas fisherianas obtienen (1) todas las posibles asignaciones; (2) el valor de un estadístico (como el efecto del ejemplo) en todas ellas; y (3) la probabilidad de obtener un valor igual o más extremo al observado bajo H_0.

Nota: En el ejemplo hemos usado la diferencia de medias (efecto) para ordenar las posibles muestras, pero cualquier estadístico puede ser usado: efecto tipificado o valor del test t de Student. En lugar de ordenar las diferencias de medias, podríamos ordenar su cociente señal/ruido (estadístico t); o, incluso, el p valor obtenido en tablas correspondiente a t.

La aplicación del principio fisheriano al reto de multiplicidad consiste en extender el recuento a todos los valores de p posibles: el p valor fisheriano estima la probabilidad de que cualquiera de los p valores calculados tenga un valor más extremo que el observado –asumiendo que son ciertas todas las hipótesis nulas puestas a prueba.

Recuerde

El p valor ajustado por el método de Fisher proporciona la probabilidad de obtener un valor más extremo asumiendo que todas las H_0 son ciertas.

Ejercicio 1.12

¿Cuántas asignaciones posibles hay si queremos repartir 30 pacientes en dos grupos de forma equilibrada? ¿Y si fueran 200? ¿Y 5000?

Si el número de casos crece, los cálculos pueden hacerse pesados, incluso para un ordenador. Una solución es obtener un número finito de sub-muestras, método conocido como 'remuestreo por Bootstrap'.

Nota: Los métodos de remuestreo se basan en los datos originales observados e incluyen, de forma natural, las relaciones entre las pruebas consideradas, por lo que obvian la necesidad de simplificar y asumir independencia.

Los métodos combinatorios fisherianos, al cambiar la asignación de los pacientes a las intervenciones, pero conservar las relaciones entre las variables, evitan la premisa de independencia entre las variables. Por eso no incurren en sobre ajustes y conservan al máximo la potencia del estudio. Sin embargo, la imposibilidad de recorrer todas las posibles asignaciones obliga a recurrir al remuestreo, por lo que diferentes ejecuciones pueden conducir a diferentes resultados, lo que obliga a especificar con detalle el método. Aquí no queda más remedio que aconsejar la consulta al estadístico profesional.

2. Monitorización. EC adaptativos

A diferencia de un estudio 'de laboratorio', en un EC la información llega de forma progresiva, lo que debe permitir mejoras. La Tabla 2.1 muestra algunas de ellas.

Según conozcamos o no la intervención asignada a los participantes, distinguiremos entre monitorización y análisis interinos.

2.1. Monitorización

El seguimiento y monitorización de un EC requiere siempre una gestión de calidad que puede hacerse de forma enmascarada al grupo de tratamiento asignado. Se trata de observar especialmente: el ritmo de reclutamiento, el nivel de adhesión al protocolo y la calidad de los datos recogidos. Esta información que llega de forma progresiva invita a hacer modificaciones al estudio.

Criterio de elegibilidad	Los criterios de inclusión pueden modificarse si: el ritmo de reclutamiento es bajo (para aumentarlo) o si la muestra es muy heterogénea (para restringir). En un caso extremo, el tratamiento sería eficaz en un único subgrupo, y se podría querer decidir continuar reclutando únicamente en ese subgrupo
Aleatorización	La razón de asignación puede oscilar en función de los beneficios mostrados por las diversas intervenciones de manera que se asigne con mayor probabilidad el tratamiento con más beneficios hasta el momento. Es más común en ensayos de Fase II donde compiten varios tratamientos.
Tratamientos	Cambiar las pautas o las dosis de administración de un fármaco o tratamiento en función de los beneficios o eventos adversos.
Tamaño muestral	En los diseños secuenciales se para o continua sin cambiar los objetivos, pero en otros diseños se puede desear un nuevo tamaño muestral según el efecto observado para mantener la potencia nominal.
Otros tratamientos	Para potenciar sinergias o evitar antagonismos.
Tiempos	Si no se alcanza el número de eventos o el de participantes en el tiempo previsto, puede aumentarse el seguimiento para conservar la potencia.
Eventos principales y secundarios	Cambios en algunas de las variables respuesta en función de los resultados. Lo más habitual es pasar de un evento simple (p.ej. muerte) a un evento compuesto (p.ej, muerte o accidente cardiovascular grave).
Análisis de datos	Pueden aparecer nuevos métodos que permitan mejorar la información recogida o su análisis.
Objetivos según resultados	En los estudios de no-inferioridad, una vez logrado el objetivo, se puede intentar demostrar la superioridad

Tabla 2.1. Adaptaciones deseables en un estudio

Nota: El soporte de empresas de investigación por contrato suele ser imprescindible. Las hay muy buenas y conviene seguir fielmente sus protocolos, pero también ser comedido y decidir recoger solo aquellas variables esenciales para el éxito del estudio, ya que incluir variables secundarias puede encarecer innecesariamente el estudio o, lo que es peor, dificultar o el cumplimiento del protocolo o la recogida de la información esencial.

Ejercicio 2.1

¿Cuáles de las adaptaciones de la tabla 2.1. son el resultado de una planificación optimista (irreal) y deberían haber sido previstas en el protocolo?

Ejercicio 2.2

¿Cuáles de las anteriores necesitan desvelar el tratamiento asignado y cuáles pueden hacerse de forma enmascarada?

Ejercicio 2.3

¿Qué consecuencias no deseadas puede tener romper el enmascaramiento de los datos pasados?

2.1. Análisis interinos

Se trata de analizar los resultados parciales del estudio para valorar si conviene tomar decisiones que alteren aspectos esenciales del mismo. Los diseños adaptativos definen en el protocolo el proceso para tomar estas decisiones. Para evitar que el conocimiento de estos resultados parciales pueda condicionar el comportamiento futuro de los investigadores y dificultar la interpretación conjunta de los resultados, conviene crear un grupo externo de expertos independientes que asesoren al comité ejecutivo.

Ejemplo 2.1: NIH creó un grupo de trabajo que publicó sus guías para la investigación cráneo-facial y dentista.

La misión del comité externo será analizar la información intermedia de eficacia y seguridad. Este apartado aborda el análisis formal de eficacia necesario para soportar estas decisiones.

Ejemplo 2.2: (Extraído de Yun-Fan, 2004) *The data and safety monitoring board consisted of three independent hepatologists, who were not members of the end-points committee, and an independent statistician. The board protected the ethical interests and safety of the patients by reviewing interim analyses. The board was empowered to recommend termination*

of the study on the basis of safety concerns or as soon as sufficient evidence indicated that lamivudine was statistically superior to placebo or that lamivudine did not provide a significant advantage over placebo. („) the study was terminated at the second interim analysis, because results had crossed the predefined boundary for showing efficacy.

Recuerde

Un comité independiente del equipo investigador, conocedor del grupo de intervención asignado, analiza eficacia y seguridad; y decide sobre la continuidad o no del estudio.

Nota: Este análisis suele centrarse en eficacia, ya que el estudio de seguridad abarca efectos generalmente imprevistos, lo que impide diseñar el estudio con control de los riesgos estadísticos. El análisis de seguridad será, por lo general, descriptivo: la simple observación de eventos no esperados ni deseados puede ser determinante para parar el estudio.

Ejercicio 2.4

¿Cuáles de las siguientes son ciertas? (1) el análisis de la calidad de los datos y del ritmo de reclutamiento no necesita desvelar el grupo de intervención; (2) el estudio de la adhesión al protocolo de intervención suele no necesitar desvelar el grupo de intervención; (3) para poder ser considerado como adaptativo, el proceso de decisión debe estar especificado en el protocolo; (4) el análisis intermedio de seguridad incluye inferencia estadística; (5) un buen protocolo recogerá el máximo posible de variables con la máxima calidad; (6) Conviene que las pequeñas oscilaciones aleatorias de eficacia y seguridad observadas en los análisis intermedios formales no alteren el comportamiento futuro de los investigadores.

2.2. Diseños adaptativos

Un ensayo clínico adaptativo es aquél que antes de iniciar el estudio planea la posibilidad de modificar, basándose en análisis intermedios formales, uno o varios aspectos del diseño –incluso sus hipótesis.

No se consideran diseños adaptativos las enmiendas al protocolo o revisiones no previstas, sea por hallazgos inesperados o por informaciones de fuentes externas.

Recuerde

Un diseño adaptativo está previsto: no requiere enmiendas.

Estos análisis deben ser realizados por un comité externo para que las evaluaciones enmascaradas no puedan introducir sesgo. Los análisis no enmascarados y no planeados de los datos, que pueden conllevar modificaciones, voluntarias o no, en la conducción del estudio, añaden incerteza a la interpretación de los resultados.

2.3. Razones para detener un ensayo

La Tabla 2.2 muestra una lista de motivos para detener un ensayo según la información proceda de monitorización enmascarada, análisis formales interinos o de fuera del estudio

Información del propio estudio		Información externa
Relacionadas con la ejecución (monitorización enmascarada)	Relacionadas con los resultados (análisis interinos)	al estudio
1. Reclutamiento inadecuado de pacientes.	1. Evidencia de diferencia en la eficacia	1. Resultados de otros estudios o meta-análisis sobre eficacia o seguridad
2. Insuficiente número de eventos que conlleven a baja potencia	2. Número o gravedad inaceptable de eventos adversos en uno de los grupos	2. Información proveniente de la práctica clínica
3. Seguimiento inadecuado (muchas pérdidas, desenmascaramiento elevado, graves desviaciones, poca adherencia a las intervenciones,…)	3. Falta de diferencias que haga improbable demostrar eficacia.	3. Cambios en la práctica clínica que hacen el estudio innecesario
4. Errores en la gestión de datos o pérdida en su calidad		4. Nuevos avances en los tratamientos
5. Falta de financiación		5. Retiro del mercado del tratamiento en estudio

Tabla 2.2. Posibles motivos para detener un ensayo. Adaptada de Muñoz et al.

Un ensayo clínico, en general, continúa hasta que haya una ventaja significativa de una de las intervenciones o bien sea poco probable que el estudio pueda demostrar diferencias. Sin embargo, también existen otras razones basadas en argumentos no estadísticos, por ejemplo, que el patrocinador vea inviable fabricar el fármaco de manera adecuada para su producción comercial; o por motivos económicos: falta de financiación, ausencia de mercado potencial, que la competencia saque al mercado un fármaco con efectos similares al pretendido…

Recuerde

Distinga entre parada temprana inesperada e interrupción programada.

Los diseños adaptativos son relativamente recientes y como tales, generan cierto escepticismo. La Tabla 2.3 resume sus ventajas e inconvenientes conocidos en su corta historia.

Ventajas	Inconvenientes
1. Eficiencia en la obtención de información. 2. Reducen el tamaño y duración de los estudios. 3. Permiten incorporar estadios exploratorios en estudios confirmatorios. 4. Mayor probabilidad de alcanzar los objetivos del estudio. 5. Mejor comprensión de los efectos del tratamiento. 6. La flexibilidad de los estudios adaptativos permite la evaluación inicial de un mayor rango de opciones. 7. Eficiente descarte de opciones sub-óptimas.	1. Riesgo de aumento del error tipo I (multiplicidad de análisis). Debe tenerse en cuenta el análisis y discusión. 2. Estimaciones del efecto sesgadas. 3. Resultados difíciles de evaluar. 4. Mayor dificultad de interpretación. 5. Posibilidad de introducir decisiones subjetivas durante el estudio (*called operational bias*), sobretodo en análisis no enmascarados conllevando sobreestimación de los resultados más favorables. El conocimiento de los grupos de tratamiento o de las diferentes adaptaciones del diseño puede influir a los investigadores. 6. Menor tiempo entre estudios para examinar detenidamente los datos entre fases y poder mejorar el diseño siguiente. 7. Las agencias reguladoras del medicamento son todavía reacias a considerar algunos de estos tipos de diseños. 8. En los diseños secuenciales, la interrupción programada por eficacia podría no aportar suficiente información sobre seguridad.

Tabla 2.3. Pros y contras de los diseños adaptativos. Adaptada de Muñoz et al.

2.1. Pasar de no inferioridad a superioridad

El objetivo de un estudio es previo a su inicio. Pero podría ser que, una vez terminado, nos demos cuenta de que podía haber sido más ambicioso.

Ejemplo 2.3: se sospecha que la pauta de la intervención es más larga de lo necesario. Diseñamos un estudio para demostrar que una intervención más corta supera a la larga en beneficios, pero al terminar el estudio, vemos que empatan.

Y decimos: qué lástima, si un empate es suficiente para adoptar una intervención más breve, deberíamos haber iniciado un estudio de no inferioridad.

Ejercicio 2.5

Las guías desaconsejan cambiar el objetivo de superioridad a no inferioridad. Repase el punto de sensibilidad en el capítulo 13 y busque argumentos para este consejo.

Ahora bien, un estudio diseñado para establecer no inferioridad garantiza, por diseño, su sensibilidad para detectar diferencias. Si luego resulta que la intervención en estudio no sólo iguala la referencia sino que incluso la mejora, el estudio demuestra 2 cosas: primero que tenía sensibilidad y segundo, que el tratamiento en estudio es superior.

Ejercicio 2.6

¿Cuál es la principal conclusión de la discusión de la agencia europea del medicamento sobre el intercambio de objetivos de no inferioridad y superioridad

Recuerde

A inicios de 2014, hay 2 ensayos adaptativos bien aceptados: diseños secuenciales y pasar de no inferioridad a uno más ambicioso de superioridad.

3. Ensayos clínicos secuenciales

El capítulo 12 mostró que el tamaño muestral de un estudio con tamaño fijo descansa en parámetros que pueden no ser bien conocidos al inicio del estudio.

Ejemplo 3.1: el efecto Δ de la intervención o la dispersión σ pueden ser distintos de los asumidos durante el cálculo de la 'n'.

Historieta: determinar el tamaño muestral es un ejemplo de ciencia-ficción.

Ejercicio 3.1

El efecto Δ de la intervención y la dispersión σ de la variable respuesta ¿forman parte de la definición del objetivo y de las hipótesis o de las premisas? ¿Qué parece más atrevido: actualizar los objetivos o las premisas?

La información contenida en un ensayo clínico se acumula a lo largo del periodo de reclutamiento –que puede ser de meses o años. Pero, en algunos casos, con cierto subgrupo inicial de participantes se podría detener el estudio si el análisis intermedio evidenciase o bien la eficacia de la intervención o bien la futilidad del estudio.

Definición

El **análisis secuencial** realiza pruebas de hipótesis por etapas.

Recuerde

Los **momentos de** los análisis intermedios dependen de la cantidad de información acumulada (número de pacientes o eventos).

Recuerde

Los **criterios** de parada están en el protocolo y son estadísticos.

En un punto anterior, se habían visto los posibles motivos para la detención de un ensayo. En el caso de los secuenciales, no podemos decir que finalicen *tempranamente* (aunque a veces se nos escape el uso del término), ya que las posibles paradas están protocoladas y no son debidas a imprevistos durante el estudio. De hecho, en este tipo de estudios, el tamaño muestral es un resultado, ya que depende de los análisis intermedios. Los ensayos secuenciales serán más cortos cuando la eficacia real de la intervención en las condiciones del estudio más se aleje de lo esperado.

Los motivos formales para detener el estudio en un análisis intermedio pueden ser:

- **Por seguridad**. Si una de las intervenciones conlleva muchos eventos adversos.
- **Por eficacia**. Si demuestra eficacia de una de las intervenciones.
- **Por futilidad**. Si los objetivos no son alcanzables.

Ejercicio 3.2

Vaya a la página principal del NEJM y busque a través de su buscador la palabra "interim". Escoja uno de los ensayos clínicos que le retorne el motor de búsqueda que tenga una antigüedad mayor de seis meses [libre acceso]

Encuentre en el artículo:

a) ¿Quién se encarga de llevar a cabo los análisis intermedios?
b) ¿Se detuvo el ensayo en un análisis intermedio?
c) Si fue así, ¿cuál fue el motivo de la detección? ¿Cuántos pacientes habían entrado hasta la fecha y cuál era el número máximo de pacientes previstos para el ensayo?

3.1. Control del riesgo α. Ajuste por multiplicidad

El principal reto de estos estudios es mantener la probabilidad α de error de tipo I deseada. Debido a los múltiples análisis intermedios, se debe ajustar. En el control de multiplicidad visto antes (p. e. Bonferroni), se hacen todas la pruebas sea cual sea el resultado de las otras, pero ahora sólo pasaremos al análisis siguiente si no se ha parado el estudio antes.

Ejemplo 3.2: Supóngase un estudio secuencial con 2 análisis (intermedio, I, y final, F), ambos con un α=0.05. Representamos por el símbolo + a "estudio con resultado positivo" y por E a "el tratamiento es realmente Efectivo".

Para obtener un resultado positivo + en el final F se tiene que (1) haber pasado el inicial I sin detectar eficacia (probabilidad de 0.95 bajo H_0); y (2) obtener + en el F (0.05 bajo H_0). Entonces el riesgo global α_G bajo H_0 es:

$\alpha_G = P(+|noE) = P(+ \text{ en } I|\ noE) + P[+ \text{ en } F|\ (- \text{ en } I \wedge noE)] = 0.05 + 0.05 \cdot 0.95 = 0.0975$

Nótese que el riesgo global es de casi el 10%, el doble de lo deseado.

Ejercicio 3.3

Calcule el riesgo global α_G suponiendo 3 análisis (2 intermedios, I_1 e I_2 y uno final, F), cada uno de ellos con un riesgo $\alpha = 0.05$

Existen varias formas para repartir este riesgo. En los ensayos secuenciales, los métodos más habituales son el de Pocock (asignación de riesgo algo mayor al inicio) ó el de O'Brien-Flemming (asignación mayor al final). Este último es más recomendable ya que concentra el riesgo cuando se dispone de mayor información, lo que preserva la potencia final del estudio.

Además de este enfoque que reparte el riesgo según la función de gasto de α, está la definición de puntos fronteras (Prueba Triangular) que además pretende poder parar el estudio si disminuyen las posibilidades de éxito.

3.2. Pruebas de Pocock y O'brien-Flemming

La función de gasto de α proporciona la probabilidad acumulada de error de Tipo I en función del tamaño muestral recolectado y permite fijar la cantidad de error que se desea gastar en cada análisis. La característica principal de esta función es que al finalizar el último análisis, esta función debe valer exactamente α (la significación deseada).

Su uso es simple, porque permite la realización de los análisis intermedios sin tener en cuenta las múltiples pruebas, únicamente considerando que habrá unos valores críticos variables en cada análisis.

Ejemplo 3.3: Supóngase un ensayo clínico con 5 análisis intermedios donde el análisis principal es una comparación bilateral Z de proporciones. Los puntos críticos para determinar la eficacia se muestran en la siguiente tabla para las metodologías de Pocock y O'Brien-Flemming.

	Pocock			**O'Brien-Fleming**		
	Valor crítico	Gasto de α	α acumulado	Valor crítico	Gasto de α	α acumulado
1r Análisis	2.41	0.016	0.016	4.23	0.000	0.000
2º análisis	2.41	0.012	0.028	2.89	0.001	0.001
3r análisis	2.41	0.009	0.037	2.30	0.007	0.008
4º análisis	2.41	0.007	0.044	1.96	0.017	0.024
5º análisis	2.41	0.006	0.050	1.74	0.026	0.050

Se rechazará la hipótesis nula en cualquiera de los análisis intermedios si el valor absoluto del estadístico de la prueba Z, (|Z|), es mayor que el valor de la tabla anterior.

Nota: En la tabla del ejemplo anterior se asume que el reparto de los participantes es equitativo entre análisis, es decir, si se han reclutado X pacientes después del primer análisis, después del segundo se habrán reclutado 2X, después del 3º, 3X y así sucesivamente.

Nota técnica: La función de gasto de Pocock viene dada por la expresión $\alpha_G \cdot (\ln(1 + (e - 1) \cdot t)$ mientras que para O'brien-Flemming, la función es $2 \cdot \left(1 - \Phi\left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{\sqrt{t}}\right)\right)$ donde t es el tiempo de realización del análisis estandarizado entre 0 y 1 y Φ es la función de distribución de la normal estándar.

Ejercicio 3.4

Un médico de familia desea comparar dos tipos de tratamientos para dejar de fumar: parches de nicotina y Vareniclina. Diseña un estudio donde el primer análisis intermedio lo realiza con los primeros 20 voluntarios (por grupo) que desean dejar de fumar. Al terminar el seguimiento, en el grupo de los parches siguen sin fumar 8 de los 20, por 16 de 20 en el de Vareniclina. Basándose en el estadístico de más abajo, y según el criterio de Pocock, ¿debe finalizar el estudio?¿Y según O'Brien-Flemming?

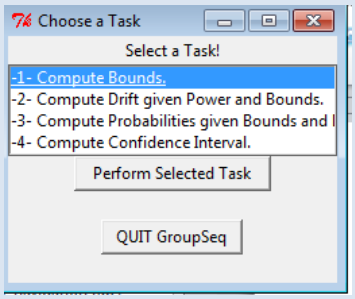
$$Z^* = \frac{p_1 - p_2}{\sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

En R, la función *groupseq* del paquete *GroupSeq* permite calcular los límites para un número determinado de análisis intermedios con una interfaz muy amigable.

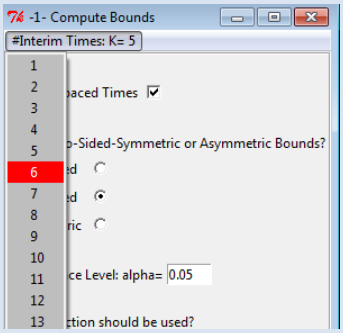
Ejemplo de R

```
# Cálculo de los límites con 'groupseq'
> install.packages(' GroupSeq')
> library(GroupSeq)
# Se abrirá una interfaz nueva (en caso contrario,
escriba
# groupseq() en la consola)
```

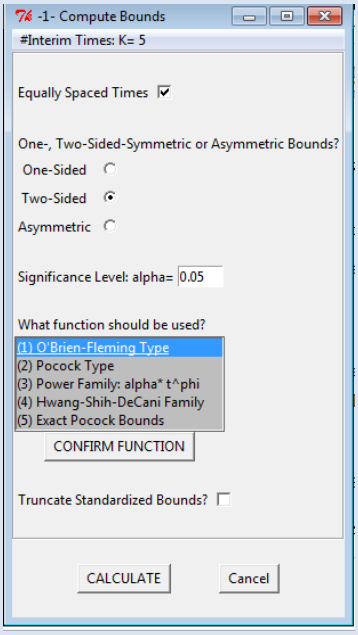
```
# Paso 1: Escoger la opción "Compute Bounds" (Calcular
# Límites) y clicar en "Perform selected Task"
```



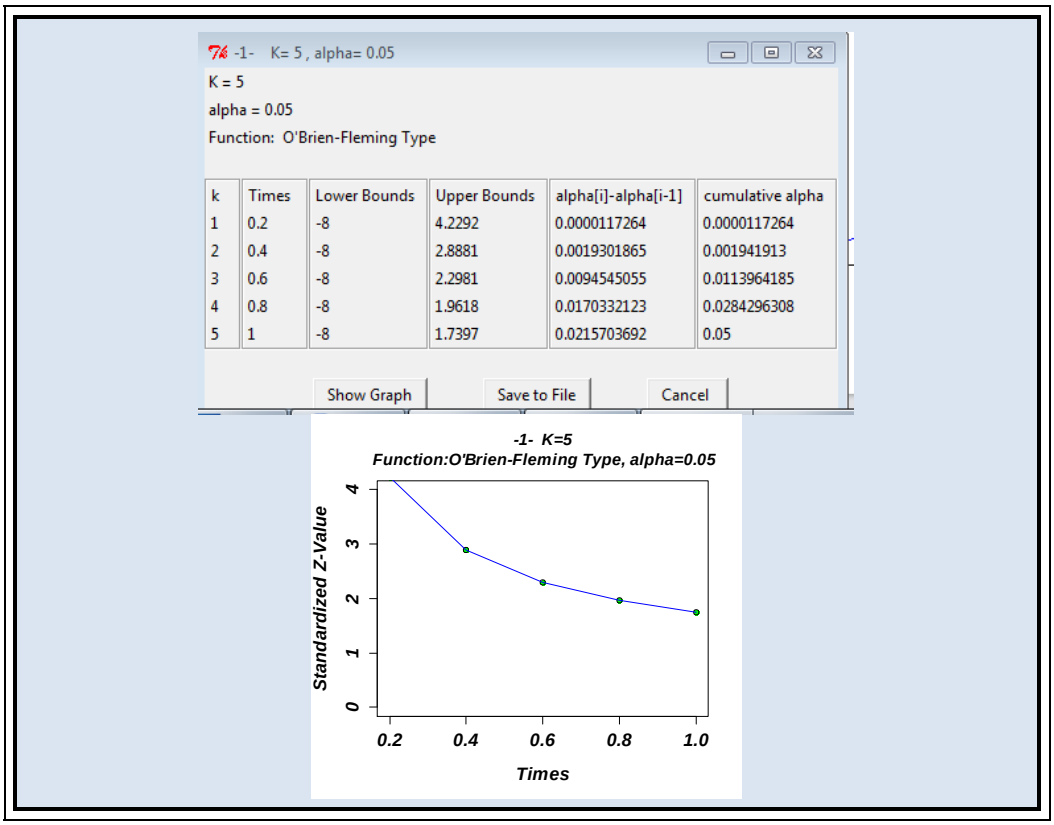

```
# Paso 2: Escoger el número (k) de análisis intermedios
# deseados sleccionandolos en la parte superior
izquierda.
```




```
# Paso 3: Escoger el resto de parámetros: (1) Análisis
equiespaciados; (2) Uni o bilateral; (3) α global; y(4)
Método.
```




```
# Paso 4: Obtengan los límites mediante "CALCULATE"
```



k	Times	Lower Bounds	Upper Bounds	alpha[i]-alpha[i-1]	cumulative alpha
1	0.2	-8	4.2292	0.0000117264	0.0000117264
2	0.4	-8	2.8881	0.0019301865	0.001941913
3	0.6	-8	2.2981	0.0094545055	0.0113964185
4	0.8	-8	1.9618	0.0170332123	0.0284296308
5	1	-8	1.7397	0.0215703692	0.05

Ejercicio 3.5

Con el paquete GroupSeq, calcule los valores de los límites para un ensayo con cuatro análisis intermedios equidistantes y con pruebas unilaterales ($\alpha = 0.025$) para Pocock y O'Brien-Flemming

3.3. Prueba triangular*

Suponga que sólo desea demostrar que el tratamiento en estudio es superior y no tiene interés en demostrar que es inferior. A cambio, quiere poder parar el estudio pronto si disminuyen las posibilidades de alcanzar el objetivo de demostrar eficacia.

La prueba triangular descansa en la razón de verosimilitudes secuencial y calcula en cada análisis los estadísticos B y V —funciones, respectivamente, de la magnitud del efecto y de la cantidad de información.

Estos estadísticos se dibujan en un plano junto con los puntos frontera: límites que indican la finalización del estudio. Estos estadísticos están definidos de forma que sean independientes entre sí.

Ejemplo 3.4: La Figura 3.1 muestra un estudio secuencial basado en la prueba triangular con 3 análisis intermedios y uno final.

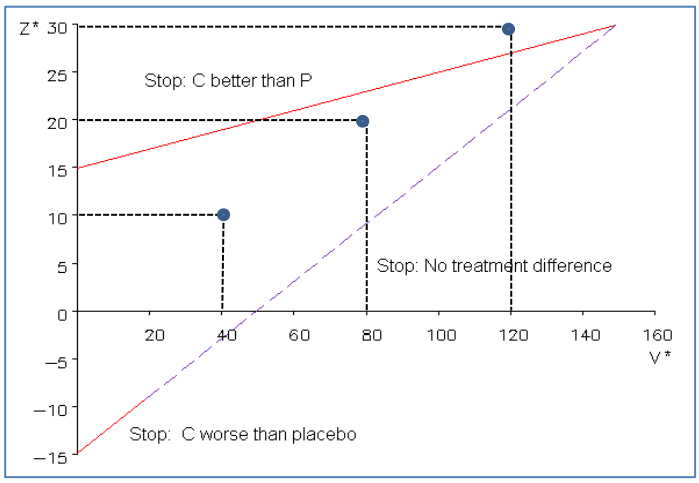


Figura 3.1. Estudio secuencial basado en la prueba triangular

Los límites para los estadísticos están representados por dos líneas (roja y lila punteada) que conjuntamente con el eje de ordenadas forman un triángulo (de ahí el nombre). Los puntos azules denotan los valores de los estadísticos B y V en los 3 primeros análisis (ver tabla siguiente).

	B	V
1r análisis	10	40
2o análisis	20	80
3r análisis	30	120

El estudio finalizó después del 3r análisis ya que los estadísticos rebasaron la frontera. Al sobrepasarla por encima, el estudio demostró el beneficio del tratamiento C respecto a P.

3.3.1. Cálculo de los estadísticos B y V

El cálculo de los estadísticos B y V del test triangular dependen del tipo de respuesta:

1) continua con distribución Normal (p.ej., la PAS);

2) dicotómica (p.ej., presencia de infección después de una intervención quirúrgica); ó

3) tiempo hasta un evento (p.ej., tiempo hasta la muerte en un estudio de supervivencia al cáncer).

La **Tabla 3.1** muestra el cálculo de los estadísticos según la respuesta.

	Tipo de respuesta						
	Normal			**Dicotómica**			**Tiempo hasta un evento**
Datos necesarios		**C**	**E**	**Total**	**C**	**E**	**Total**	HR: Hazard Rate Ratio
	Media	Y_{iC}	Y_{iE}		**Éxito**	S_{iC}	S_{iE}	S_i	
	Variancia	S_{iC}^2	S_{iE}^2		**Fracaso**	F_{iC}	F_{iE}	F_i	
	Tamaño	n_{iC}	n_{iE}	n_i		n_{iC}	n_{iE}	n_i	
						$P_C = S_{iC} / n_{iC}$		$P_E = S_{iE} / n_{iE}$	
Efecto (θ)	$\theta = \mu_E - \mu_C$				$\theta = \ln\left\{\frac{P_E(1-P_C)}{P_C(1-P_E)}\right\}$			θ = Ln(HR)
B_i	$B_i = \left\{\frac{n_{iC}n_{iE}}{n_i S^2}\right\}^{1/2} \phi^{-1}\{T_i(t_i)\}$				$Bi = \frac{n_{iC}S_{iE} - n_{iE}S_{iC}}{n_i}$			B_i = estadístico LogRank
V_i	$V_i = \frac{n_{iC}n_{iE}}{n_i S^2}$				$Vi = \frac{n_{iC}n_{iE}S_iF_i}{n_i^3}$			$V_i \approx$ #events / 4

Tabla 3.1. Calculo de los estadísticos en el test triangular

3.3.2. Reglas de decisión

Cada análisis intermedio valora si el estadístico supera la frontera para tomar la decisión.

En el primer análisis intermedio, por ejemplo…

- Se concluye que E es más eficaz que C si $B_1 \geq U_1\sqrt{V_1}$
- Se concluye que no se podría demostrar que E sea más eficaz que C si $B_1 \leq L_1\sqrt{V_1}$
- Se continua el estudio si $B_1 \in (L_1\sqrt{V_1}, U_1\sqrt{V_1})$

En el último análisis (K-ésimo), sólo existen dos opciones:

- Se concluye que E es más eficaz que C si $B_k \geq U_k\sqrt{V_k}$
- No hay evidencia que E sea más eficaz que C si $B_k \leq U_k\sqrt{V_k}$

El diseño de ensayos secuenciales debe calcular V (que determinará en qué momento se "mira" el ensayo) y los límites L_i, U_i de B_i para cada parada.

3.3.3. Caso de diseño con 2 análisis

Para poder definir los criterios de parada, se tienen 5 parámetros desconocidos: L_1, U_1, U_2, V_1, V_2. Para hallar el valor de estos 5 parámetros se necesitan 5 ecuaciones o, dicho de otra manera, 5 restricciones. Sin embargo sólo hay 2; las correspondientes a la imposición de los riesgos α y β. Se deben añadir 3 restricciones adicionales para poder hallar el valor de todos los parámetros. Algunas restricciones razonables son:

1) $\boldsymbol{V_2 = r \cdot V_1}$. Siendo r = 2 si el tamaño del análisis final es el doble del análisis intermedio
2) $\boldsymbol{L_1 = c \cdot U_1}$. Siendo c = -1 si se utiliza una regla simétrica (misma probabilidad de demostrar eficacia de un tratamiento u otro) o c = 0 si se para por futilidad – es improbable encontrar evidencia de que E > C en análisis posteriores.
3) $\boldsymbol{U_2 = d \cdot U_1}$. Siendo d = 1 ó d = min(V_2) ó d = min($E[V^*|\theta_0]$) ó d = min($E[V^*|\theta_A]$)

En diseños con más paradas hay que ampliar el número de restricciones adicionales.

Ejercicio 3.6

Ojee el artículo de Bolland et al. sobre el análisis de un diseño secuencial aplicado al estudio ICTUS y conteste las siguientes cuestiones:

a) [Primer párrafo en pág. 140]. ¿Cuál era la potencia y el valor de α para este estudio?

b) [Primer párrafo en pág. 141]. ¿Cuál hubiese sido el tamaño muestral del estudio si se hubiese realizado con un tamaño fijo?

c) [Penúltimo párrafo, pág. 141]. ¿Cuántos análisis intermedios se planearon? ¿Con cuántos pacientes?

d) [Penúltimo párrafo, pág. 141]. ¿Cuáles fueron los límites críticos superiores e inferiores en estos análisis?

e) [Primer párrafo, pág. 142]. ¿Cuál era el riesgo α acumulado en cada análisis?

f) Compare los anteriores riesgos con los asumidos con el método de O'brien-Flemming en las mismas características usando la función *groupseq*. [Nota dado que los tiempos no son equidistantes, contando que en el primer análisis hay más pacientes, estos se han de especificar como proporcionales al tamaño pero en escala de 0 a 1: 0.385 (1000), 0.590 (1533), 0.795 (2067), 1 (2600)]

g) [Penúltimo párrafo, pág. 142] ¿Cuál seria la probabilidad de llegar al último análisis (n = 2600) si no hubiera efecto del tratamiento?

3.4. Controversia sobre los diseños secuenciales*

Se ha afirmado que un interés de los estudios secuenciales es parar el estudio *tempranamente* para que el promotor pueda ahorrarse los costes de introducir más pacientes. Sin discutir si este objetivo es o no lícito, hay que resaltar que parar *en el momento adecuado* permitirá emplear la mejor opción terapéutica en un mayor número de pacientes. Así, si aceptamos este último objetivo, la discusión técnica debe ser si el diseño secuencial (1) controla adecuadamente los riesgos de decisiones erróneas; y (2) la estimación del efecto que proporciona es insesgada.

La estimación del efecto del tratamiento en los ensayos que finalizan de forma temprana mostrando beneficio de alguna de las intervenciones, está sesgada en el sentido de que magnifica el efecto de la intervención.

Lectura: Bassler et al comparan las estimaciones de estudios que han finalizado tempranamente con estudios similares de muestra fija o que hubiesen completado todos los análisis posibles:

"*Study Selection: Selected studies were RCTs reported as having stopped early for benefit and matching nontruncated RCTs from systematic reviews" (...) "Truncated RCTs were associated with greater effect sizes than RCTs not stopped early*".

Nótese que el diseño es muy discutible ya que comparan estudios significativos finalizados tempranamente con todos los estudios (significativos o no) que hicieron un único análisis final.

La Figura 3.2 proporciona una explicación no formal de este sesgo. El gráfico de la izquierda parte de la hipótesis de ausencia de efecto del tratamiento representado por un punto azul. Los puntos negros representan una simulación de los efectos de 100 estudios que se hubiesen hallado en los 2 análisis intermedios y en el final. Las líneas rojas discontinuas marcan el límite a partir del cual se pararía el estudio y las líneas verdes representan el efecto esperado para los estudios que finalizan en un determinado instante.

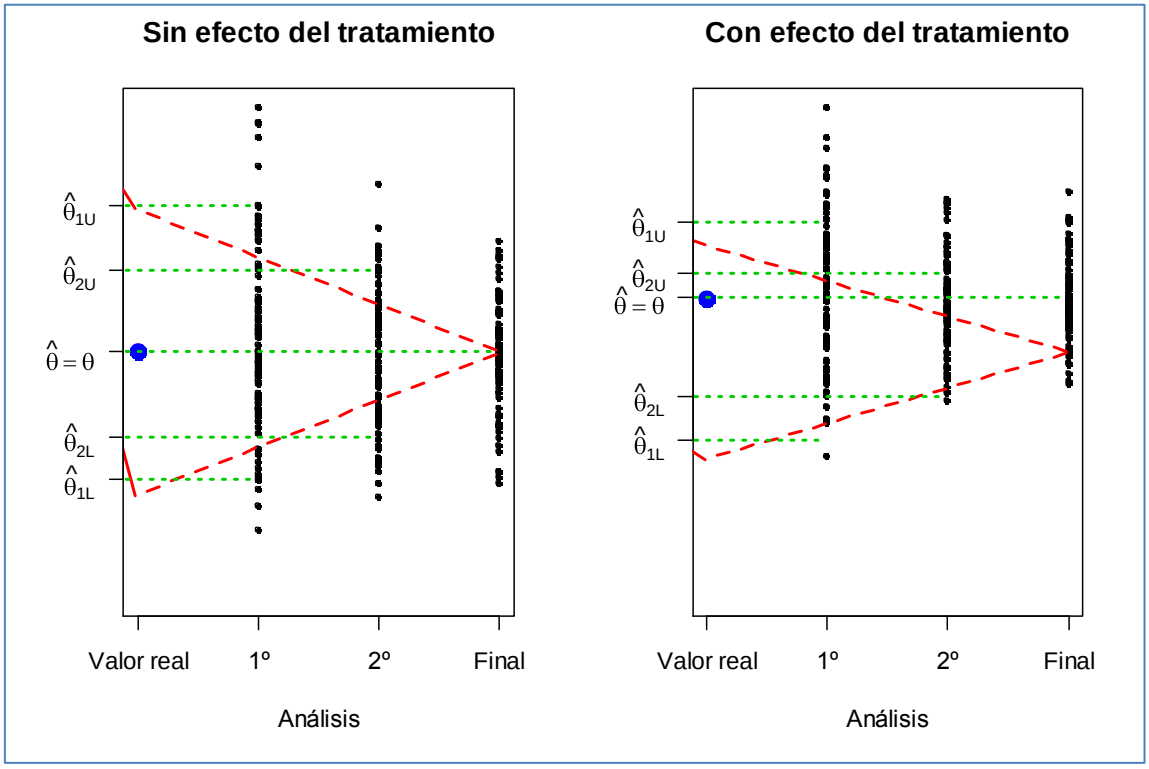


Figura 3.2. Sesgo en función del tamaño del estudio

Obsérvese, por ejemplo, que para el primer análisis, el promedio de los 100 efectos coincide con el valor real, pero si nos centramos únicamente en los resultados significativos de la parte superior, su promedio es muy superior al valor real. Esta es la explicación del sesgo. El valor esperado del efecto en el primer análisis intermedio coincide con el valor real del efecto, pero el valor esperado del efecto en el primer análisis condicionado a que se ha parado por eficacia, no coincide en absoluto.

Desgraciadamente, Stephen Senn explica que este sesgo aplica de forma más genérica a todo tipo de ensayos. Un diseño insesgado es aquel en el que el promedio de todos los resultados posibles coincide con el valor del auténtico parámetro de interés. Pero eso no implica que el promedio de todos los parámetros coincida con la estimación obtenida. Posiblemente, se trata de una versión moderna del problema de regresión a la media.

Soluciones a los ejercicios

1.1 Desgraciadamente, en muchos EC la respuesta es que sí, que puede pasar. Pero si es un pivote del que dependen acciones posteriores, el criterio para tomar la decisión debe estar perfectamente definido en el protocolo, así como los métodos para controlar los riesgos estadísticos.

1.2 Sí. Por ejemplo, una variable podría ser principal, concentrando los riesgos, y la otra secundaria, con valor para ratificar que, si los resultados se solapan razonablemente, un análisis de sensibilidad de las conclusiones a las elecciones del diseño confirma que otras elecciones llevan a conclusiones similares. También, haber especificado ambas como principales y que era preciso que ambas debían ser significativas. Pero si hubieran dicho que el resultado sería positivo si cualquiera fuera positiva, entonces, habría que ajustar –y perder la significación.

1.3 La correcta es la b: "Sensitivity analyses are useful to investigate whether or not the main results are consistent with those obtained with alternative analysis strategies or assumptions". Si encuentra en las guías algo que apoye la afirmación 'a', les rogamos que nos informe.

1.4 Menciona todas ellas como premisas en las que descansa un estudio observacional. Un buen análisis de sensibilidad debería descartar que las conclusiones dependan de alguna de ellas.

1.5 0.05/10=0.005. Al menos una prueba debería ser significativa al 0.5% para que el estudio fuera positivo.

1.6 $1-(1-0{,}05)^{0.1} = 0.0051162$.

b) Sí, al pedir un riesgo α más pequeño, las probabilidades de alcanzar resultados positivos disminuyen: se pierde potencia.

1.7 Siguiendo el proceso, al poner a prueba H_{03}, debe controlar que hasta k=3 hipótesis nulas podrían ser ciertas, por lo que $P = 0.0137 < \alpha_3 = 0.05/3 = 0.0167$, se rechaza H_{03}. Pero al poner a prueba H_{04} $P = 0.0324 > \alpha_4 = 0.05/2 = 0.025$, nada se opone a aceptar H_{04} y se para el proceso. En resumen, rechazamos las 3 primeras y aceptamos las 2 últimas.

1.8 Siguiendo el proceso de Hockberg, al poner a prueba el tercero hay que tener en cuenta que 3 podrían ser simultáneamente ciertos, por lo que $P = 0.0137 < \alpha_3 = 0.05/3 = 0.0167 >$, se rechaza y se para el proceso, llegando a la misma conclusión anterior.

1.9 Las pruebas de Bonferroni y Sidak ajustan por 5 posibles hipótesis nulas siempre, pero los otros 2 solo por las que quedan por rechazar (Holm) o las ya no rechazadas (Hochberg), 3 en ambos casos: 0.014*3≈0.041.

1.10

(a) Ajuste por Bonferroni:

Al poner a prueba las diferentes hipótesis, se considera un $\alpha_i = 0.05/7 = 0.00714$:

Observamos que todos los P valores obtenidos, excepto el obtenido para HDL, son mayores que el nivel de significación individual ajustado, por consiguiente, se rechaza la hipótesis nula de HDL y se aceptan el resto.

(b) Ajuste por Sidak:

Al poner a prueba las diferentes hipótesis, se considera un $\alpha_i = 1 - (1 - 0.05)\text{^}(1/7) = 0.0073$.

Mismas conclusiones que con el método de Bonferroni.

(c) Método de Holm (+Bonferoni):

Ponemos a prueba las diferentes pruebas de hipótesis en orden creciente, según el valor de P:

Al poner a prueba la variable HDL, hay que tener en cuenta que hasta k=7 hipótesis nulas podrían ser ciertas, por lo queP $= 0.006 < \alpha_1 = 0.05/7 = 0.00714$, se rechaza.

Al poner a prueba la variable T, hay que tener en cuenta que hasta k=6 hipótesis nulas podrían ser ciertas, por lo que $P = 0.08 > \alpha_2 = 0.05/6 = 0.00833$, se acepta.

Al aceptar la variable T, y teniendo que el resto de variables tienen un P valor superior, se aceptan el resto de hipótesis nulas.

(d) Método de Hochberg (+ Bonferroni):

Ponemos a prueba las diferentes pruebas de hipótesis en orden descendiente, según el valor del P valor:

La primera variable que ponemos a prueba es FC, con un P valor claramente superior a 0.05, se acepta.

Al poner a prueba la variable LDL, se tiene que tener en cuenta que podrían haber 2 simultáneamente ciertas, $p = 0.314 > \alpha_2 = 0.05/2 = 0.025$, se acepta.

La siguiente en ponerse a prueba es CT, con un $p = 0.123 > 0.05/3 = 0.0167$, se acepta.

La variable PAS es la primera variable en ser rechazada, ya que $P = 0.012 < 0.05/4 = 0.0125$.

Por consiguiente, también se rechazan las hipótesis nulas referidas a las variables PAD, HDL y T, al tener un p valor inferior al de la variable PAS.

Ejercicio reto. Como todas las comparaciones son entre el grupo 0 con 3 casos y los 3 casos del otro grupo, el error típico es: 0.617*raíz(2/3)≈0.504, por lo que los 6 t-test valen -0.040, -0,119, 2.461, 2.382, 2.203 y 2.223. Empezamos por poner a prueba la dosis de 6 g y sólo seguimos si fuera significativa (para proteger el α global): Rechazamos todas excepto las de 1 y 2 g.

1.11 Se quieren asignar 8 pacientes en dos grupos de 4 pacientes cada uno.

```
> choose(8,4)
[1] 70
```

Hay 70 combinaciones de 8 pacientes tomados de 4 en 4.

La probabilidad que la combinación elegida haya sido TTTTCCCC es de 1/70=0.01428.

1.12 Si se quieren asignar 30 pacientes en dos grupos de 15 cada uno:

```
> choose(30,15)
[1] 155117520
```

Con sólo 30 pacientes, deberíamos calcular el resultados para más de 150 millones de posibles asignaciones. Un buen reto, accesible sólo para buenos ordenadores bien programados.

Si se quieren asignar 200 pacientes en dos grupos de 100 cada uno:

```
> choose(200,100)
```

[1] 9.054851e+58

Con 200 pacientes, el resultado tiene casi 60 cifras antes del punto decimal. Un reto incluso para el Mare Nostrum de la UPC.

Y, si se quieren asignar 5000 pacientes en dos grupos de 2500:

> choose(5000,2500)

[1] Inf

Observe como el número de combinaciones posibles es tan alto que R da cómo respuesta infinito.

2.1. Los puntos que deberían de haberse previsto de forma más efectiva en el protocolo son, por lo menos:

(1) El criterio de elegibilidad. El objetivo de añadir criterios de elegibilidad es definir una muestra en el que el efecto de la intervención sea homogéneo. A los investigadores les gusta añadir muchos criterios de entrada y esto provoca que se disponga de pocos pacientes.

(2) El tiempo de recolección de los datos.

(3) Los eventos primarios y secundarios.

2.2. Los procesos en los que es necesario desvelar el tratamiento asignado son:

(1) Proceso de aleatorización

(2) Regímenes de tratamientos

(3) Tamaño muestral

(4) Introducción de tratamientos concomitantes

2.3. El conocimiento de la intervención a realizar puede influir en la actitud del responsable de administrar el tratamiento, del sujeto experimental que recibe la intervención o del analista que procesa la información resultante de la intervención. Este fenómeno puede llevar al error sistemático o sesgo.

2.4. Son ciertas todas excepto la (4) y la (5). La (4) porque el análisis intermedio de seguridad no requiere realizar inferencia y la (5) porque es una barbaridad.

2.5. Como los estudios de superioridad si salen positivos no necesitan probar la sensibilidad del estudio (capacidad para demostrar que, caso de que hubieran diferencias, el estudio hubiera podido establecerlas), al diseñarlos no se deja establecida su sensibilidad. Por tanto, si un estudio no logra demostrar superioridad, no puede argumentarse que podría establecer equivalencia o no inferioridad al no poder garantizar su sensibilidad.

2.6. Que la interpretación del IC no conlleva tantas dificultades.

3.1 Δ define la hipótesis alternativa y forma parte, por tanto, de los objetivos del estudio; pero σ, de las premisas. Por supuesto, cambiar los objetivos del estudio es mucho más comprometido.

3.2 a) Normalmente el análisis lo realiza un comité independiente

b) El ensayo puede haber finalizado antes o no de lo previsto

c) El motivo para la finalización del ensayo puede ser eficacia, futilidad o seguridad. Observe el porcentaje de participantes que se ahorraron por hacer un diseño secuencial.

3.3 α_G = P(+|noE) = P(+ en I_1| noE) + P(+ en I_2| noE) + P(+ en F| noE) = 0.05 + 0.95·0.05 + 0.95·0.95·0.05 = 0.143

3.4 $p_1 = 8/20 = 0.4$; $p_2 = 16/20 = 0.8$; $p = (p_1 + p_2)/2 = 0.6$

$Z=(0.4 - 0.8)/\sqrt{0.6 \cdot 0.4 \cdot (1/20+1/20)} = -2.58$

Con el criterio de *Pocock* se pararía el estudio ya que |Z| = |-2.58| = 2.58 > 2.41. Con el criterio de *O'Brien-Flemming* no se pararía ya que |Z| = |-2.58| = 2.58 < 4.23

3.5 Pocock: 2.36 en todos los análisis; O'Brien: 4.3326, 2.9631, 2.359 y 2.01.

3.6 a) α = 0.05; potencia = 0.80 ; b) n =2421 ; c) 4 análisis con n's = 1000, 1533, 2067 y 2600; d) UL =25.28 en todos los análisis y LL = -5.83, 4.54, 14.90 y 25.28; e) α_i = 0.0006, 0.0046, 0.0136 y 0.025; f) α_i = 0.0001, 0.0023, 0.0101 y 0.025; g) 0.0749